引用格式：鄂维南. AI助力打造科学研究新范式. 中国科学院院刊, 2024, 39(1): 10-16, doi: 10.16418/j.issn.1000-3045.20231224001.
E W N. AI helps to establish a new paradigm for scientific research. Bulletin of Chinese Academy of Sciences, 2024, 39(1): 10-16, doi: 10.16418/j.issn.1000-3045.20231224001. (in Chinese)

# AI助力打造科学研究新范式*

鄂维南

1 北京大学　北京　100871
2 北京科学智能研究院　北京　100084

**摘要**　科学研究的目的是发现基本原理和解决实际问题。尽管人类在发现基本原理和解决实际问题上已经取得了巨大成就，但有效工具和有效科研组织模式的缺乏仍然是制约科研效率的主要瓶颈。人工智能（AI）的迅速发展为改变这种状况提供了新的可能。近年来，深度学习方法在科学研究领域大放异彩，不仅助力解决了一些核心科学问题，扩展了科学方法，也开始带动科学研究从传统的“作坊模式”转向“平台模式”。目前，我国已在人工智能驱动的科学（AI for Science）领域打下良好基础，应把握机遇，争取引领科技创新，为人类的科技发展作出贡献。

**关键词**　人工智能驱动的科学研究，科学计算，安卓模式

DOI　10.16418/j.issn.1000-3045.20231224001
CSTR　32128.14.CASbulletin.20231224001

科学研究有2个主要目的：发现基本原理，如发现行星运动规律和量子力学原理；解决实际问题，如解决工程和工业中出现的问题。科学研究有2种主要方法：开普勒范式，即数据驱动的方法；牛顿范式，即基本原理驱动的方法。前者最好的例子是行星运动三定律的发现，即开普勒通过分析观察数据发现了这些规律。后者最好的例子是牛顿对行星运动三定律的解释和运用。牛顿提出了力学第二定律和万有引力定律，在此基础上将行星运动问题归结为一个常微分方程问题并推导出行星运动三定律。这里原始的科学发现是开普勒做出的，但他并不理解其背后的原因。牛顿进一步发现了背后的基本原理，这些原理进而可用于许多其他问题。

从实际应用的角度来看，在量子力学建立之后，寻找基本原理的任务已经基本完成。早在1929年，狄拉克[1]就宣称，“大部分物理学和整个化学的数学理论

*本文借鉴了笔者2023年12月1日发表在*SIAM News*上的论文*AI for Science*的部分内容，并在此基础上修改、扩写而成。

修改稿收到日期：2023年12月27日

所需要的基本物理定律已经完全被人们所知，困难在于这些定律的精确应用导致方程过于复杂而无法求解”。他的断言不仅适用于化学，也适用于生物学、材料科学，以及所有其他不涉及高能物理的自然科学与工程学科。在实际情况中，通常不必深入到量子力学层面，而可以使用一些简化的基本原理，如气体动力学的欧拉方程和流体力学的纳维—斯托克斯方程。

对于应用数学家来说，一方面有了这些基本原理，所有的自然科学和相关的工程问题都可以归结为数学问题，再具体而言是常微分方程或偏微分方程问题。另一方面，在开发出有效的工具之前，为了解决实际问题，科学家只能大幅度简化或彻底忽略这些基本原理。

冯·诺伊曼认识到计算机和数值算法应该提供一种利用这些基本原理解决实际问题的通用方法，这是一个重大进展。沿着这个方向，人们提出了许多求解这些微分方程的数值算法，如有限差分、有限元和谱方法。这些算法的基本出发点是一般函数可以用多项式或分片多项式逼近。这些工作的影响是巨大的。今天，科学计算已经成为现代技术和工程科学的基础。许多学科，如结构力学、流体力学和电磁学，由于引入数值算法而发生了彻底改变。

## 1 科学研究的基本问题

目前，科学研究中并非所有问题都得到了解决。例如研究材料的性能和设计、药物设计、内燃机设计，以及许多控制问题仍然远远做不到使用基本原理来解决。在这些领域，理论工作往往与现实世界相去甚远，现实世界的问题必须通过试错或靠经验来解决。这导致科学研究效率低下，相关领域的技术提升进展缓慢。

所有这些“困难”问题都有一个共同特点，即它们依赖于多个独立变量。所以，这些困难实际来自维度灾难。以量子力学的薛定谔方程为例，忽略对称性，波函数中独立变量的个数是粒子数量的3倍，所以10个电子的系统虽然是非常简单的体系，但其对应的30维空间偏微分方程却已经非常复杂！

## 2 人工智能为科学计算提供新的解决方法

深度学习在图像分类、图像生成和围棋等方面取得了极大的成功。这些都是标准的人工智能问题，但从数学角度来看，这些问题其实是函数逼近、概率分布的逼近和采样，以及求解贝尔曼方程的问题。而所有这些都是应用数学，尤其是计算数学长期面临的典型问题。不同之处在于，这些人工智能问题比应用数学中处理的问题维度要高得多。以图像分类问题为例，这里的自变量是图像，每个像素都是1个自由度。因此，1张32×32像素的彩色图片有3 072个自由度。换句话说，这个问题的维度是3 072。

深度学习在这些高维问题上取得的成功提示深度神经网络可能是逼近高维函数更有效的工具。虽然目前还没有建立起一个完整的深度学习的数学理论，但已经取得了一些重要进展和直观了解。首先，神经网络就是一类特殊的函数。如果使用规则网格上的分片线性函数来逼近一个函数，其误差与网格大小的平方成正比。这正是维度灾难的根源：随着维度的增加，同样网格大小所需要的格点个数呈指数增长。不仅基于分片线性函数的逼近是这样，所有基于固定基函数的逼近方法都是这样。如果利用神经网络函数来逼近一般的函数，那么至少在某些情况下，可以证明其逼近精度不会随着维度的增加而恶化，就跟计算数值积分的蒙特卡罗（Monte Carlo）方法一样[2]。

这个观察结果有着广泛的意义。因为函数是最基本的数学对象之一，所以一个新的高维函数逼近工具将对许多不同的领域产生深远影响。特别是，深度学习应该有助于解决之前讨论过的那些受维度灾难困扰的问题。这是人工智能驱动的科学（AI for Science）的出发点。

这方面最成功的例子是预测蛋白质结构的AlphaFold算法。蛋白质结构是生物学最基本的问题之一。研究蛋白质结构的基本方法是首先最小化整个蛋白质—溶剂系统的总势能。但2个主要的困难限制了这种方法的成功：获得精度足够高的势能函数，以及该函数景观的复杂性。科学家也曾尝试过数据驱动的方法，但其成功仅限于预测二级结构，如α-螺旋和β-折叠。通过充分利用蛋白质序列数据集及最先进的深度学习模型，DeepMind公司开发了AlphaFold2算法，它以非常优雅的方式基本解决了蛋白质结构问题[3]。这项研究震惊了世界。

AlphaFold2是纯粹数据驱动的方法。但这并不意味着AI for Science是一个纯粹数据驱动的研究范式。事实上，科学研究遵循如前所述的基本原理或第一性原理，而AI for Science的一个主要组成部分是用人工智能方法为这些基本原理开发更高效的算法或近似模型。在这方面，最著名的例子是分子动力学。分子动力学是生物学、材料科学和化学的基本工具，其思想是通过计算体系中原子的动态轨迹来研究分子和材料的性质。原子运动遵循牛顿定律，困难的部分来自于模拟原子之间的相互作用力或势能函数。经验势函数的方法是尽可能地猜出原子间势能函数的函数形式，然后用一些实验或第一性原理计算出的数据来拟合其中的参数。虽然这种方法可以提供一些帮助，但作为一个研究特定体系的定量工具，它是不可靠的。1985年，Car和Parrinello开发了第1个基于第一性原理的人工智能方法：通过使用量子力学模型（如密度泛函理论）来实时计算原子之间的作用力。这种方法能够以第一性原理的精度来模拟特定体系。但在实践中，效率是一个瓶颈。由于效率的限制，只能用这种方法来处理含数千个原子的体系。

机器学习提出了一种新的范式。在这个新的范式下，量子力学仅用于提供数据。基于这些数据，可以使用机器学习方法来得出原子间势能函数的精确近似，然后就像使用经验势能函数一样将其用于分子动力学模拟。

为了使这个策略真正有效，必须处理2个重要问题。① 网络架构。它应该是可拓展的，并且遵循物理学基本规律。可拓展性能够在小体系上做机器学习并将结果应用于更大的体系。这个问题在Behler和Parrinello两位科学家的经典工作中得到了解决。遵循物理规律意味着必须保持对称性、守恒律、不变性和其他物理约束。在势能函数这个问题中，需要考虑的主要是平移、旋转和置换不变性。这可以通过使用一个嵌入网络来实现，该网络将原子位置的信息映射到一组保持对称性的函数上[4]。然后再通过一个逼近网络来拟合势能函数。② 数据有关。一方面，如果希望机器学习方法产生的势能函数在所有感兴趣的实际场景中都与原始的量子力学模型一样精确可靠，那么训练数据集就需要能够对所有这些不同场景都具有充分的代表性。另一方面，由于标注数据是用量子力学模型计算出来的，而这些计算是比较昂贵的，所以希望数据集尽可能小。这就需要一种自适应数据生成算法，它能够帮助人工智能在学习过程中动态生成“最优”数据集。

ELT算法就是为了解决这个问题[5]。它由探索（exploration）、标注（labeling）和训练（training）3个部分组成，因此得名ELT。ELT可以从没有数据和粗糙的初始势能函数开始。在探索过程中，使用一些采样算法（如某种分子动力学方法）来探索不同的原子构象。对于遇到的每个构象，可以计算出一个指标值来查看是否需要对其进行标注。然后将标注好的数据添加到训练数据集中，并基于它定期更新对势能函数的逼近。

该算法的关键在于采样方案和如何计算指标值。采样方案的基本思想是仅探索实际感兴趣且缺乏足够多的训练数据的构象空间。指标值的关键在于判别哪些构象附近还缺乏足够多的训练数据。对于后者，

ELT方案采用的方案是训练一组近似势能函数。这组近似势能函数之间的标准差定义为指示函数。对当前采样到的构象，如果其指示函数值超过了阈值，就对该构象作标注。其背后的逻辑是，如果这个构象附近有足够多的训练数据，那么不同网络预测的势能函数值都应该非常准确且彼此接近。大的标准差表明附近没有足够多的训练数据，因此应该对当前构象进行标注并加到训练数据集中。对于采样算法，选择带偏差的分子动力学，其中偏差势函数由当前对势能函数的逼近来定义，并由其准确性的置信区间大小来定义权重。其背后的逻辑是，如果当前已经得到的势能函数在一个区域范围足够准确，那么应该离开这个区域而到其他地方进行采样[6]。

有了这些主要组件，确实可以为一大类（如果不是全部的话）原子体系提供具有第一性原理精度的势能函数。所得的模型称为深度势能分子动力学（deep potential molecular dynamics，DeePMD）。它是一个可靠的、具有第一性原理精度的原子模拟工具。结合高性能计算，它将以第一性原理精度分子动力学模拟的能力从只能处理数千个原子的体系扩展到处理170亿个原子的体系[7]。DeePMD软件包DeePMD-kit也大大降低了DeePMD的使用门槛[8]。

类似的想法可以应用于其他物理模型。例如，可以用高度准确的量子化学计算数据来训练更通用、更准确的密度泛函模型。还可以开发更准确、更可靠的粗粒化分子动力学模型，以及更准确的动力学方程的矩阵模型等。事实上，机器学习正是过去多尺度、多物理建模所缺少的工具[6]。

除了基本原理的模型之外，人工智能方法还可以提供更高效、更准确的反演算法，从而增强实验表征能力。先前讨论过的基于人工智能的算法可以为正问题提供更逼真、更准确的数据，而神经网络中的可微分结构可以帮助设计解决反问题的优化或采样算法。这项工作仍处于早期阶段，但它是一个有巨大发展空间的方向。

人工智能方法还有可能改变人们利用文献和现有科学知识的方式。文献和现有科学知识是科研灵感的主要来源之一。然而，利用好这些资源也是一个非常艰巨的任务：需要从大量信息中挖掘出相关文献和知识，并需要花大量时间来阅读和研究它们。然而，可以利用人工智能数据库和大语言模型来收集和整合这些信息并更有效地查询这些信息。原则上，对于感兴趣的任何研究课题，都可以使用人工智能工具快速总结文献中的相关信息及其来源。人工智能技术甚至可以帮助建议一些进一步的研究方向。这将大大提高科学研究的效率。

随着这些新的可能性的出现，可以探索一种新的科研范式，并把它称为科学研究的“安卓范式”。在这个新范式下，科学界将共同努力建立起一套新的基础设施，包括用于基本原理的人工智能算法、人工智能赋能的实验设施和新的知识数据库。这些平台构成了科学研究的“安卓平台”。无论是寻找特定化学反应中的催化剂还是设计新电池，这些针对特定应用的研究工作都可以在这个“安卓平台”上进行。这无疑将加快科学研究的进程。

这种横向整合的观点也将有助于打破学科壁垒，加强跨学科的研究和教育。横向整合的观点本身并不新，由于缺乏有效的工具，过去它难以带来实质性的进展。如前所述，人工智能方法提供了大大改进这些横向工具的空间。这些新的横向工具，例如新的查阅文献和现有科研数据的平台，以及自动化、智能化的实验平台，使得科研人员能够从横向的角度更有效地看待不同的科研场景。例如，对原子体系，生物学关注的是生物大分子，材料科学关注凝聚态体系；化学比较关注小分子，化工领域则比较关注高分子。而从理论工具的角度来说，无论哪种体系，都离不开电子结构方法和分子动力学方法。实验工具则包括不同尺度的光谱和显微镜成像技术。尽管不同领域关注不同

体系，这些不同领域的工具和知识都应该可以最大程度地共享。在这个框架下，学科之间的界限也就自然消失。

## 3 我国AI for Science的发展现状

带着这一愿景，笔者团队在2018年启动了DeepModeling开源平台①。这个平台的目的是邀请科学界共同努力，为物理建模和数据分析建立基于人工智能方法的基础设施。到目前为止，它已经产生了巨大的影响力并吸引了许多的开发者，在中国，AI for Science的发展呈现出令人欣慰的良好局面。所有这些，都为AI for Science在中国的发展奠定了良好的基础。

（1）在短短几年内，AI for Science的重要性和它带来的巨大发展空间已经得到了广泛的认可。一大批各个领域的领军学者都高度重视AI for Science这一机会。2024年初《中国科学院院刊》策划组织“大力推进科研范式变革”专题，就是一个例证。

（2）一批专注于AI for Science的研究团队正在出现并展示出良好的势头。经过3年多的酝酿，北京科学智能研究院于2021年9月在北京市的支持下正式成立。这是国际上第1个专注于AI for Science的研究机构，致力于打造AI for Science时代的基础设施。除此之外，还有中国科学技术大学的机器化学家团队、厦门大学嘉庚创新实验室的AI for Electrochemistry团队等。

（3）一批企业也在AI for Science方向积极布局。这体现了产业界对AI for Science的巨大信心。在AI for Science的旗帜下聚集了一大批有能力、有决心、有干劲的青年产业人员。

（4）科学技术部、国家自然科学基金委员会等国家机构和北京市、上海市等地方政府都在积极出台政策，支持AI for Science的研究。2022年，国家自然科学基金委员会交叉科学部首先推出“可解释、可通用的下一代人工智能重大研究计划”，AI for Science是其中一个重要组成部分。

## 4 建议

如今的良好基础并不代表AI for Science在中国的健康发展已经板上钉钉。对一个领域的发展来说，成为热点是一把双刃剑。越是热点，就越容易产生泡沫。如何才能保证利用好这个机会，让AI for Science带动我国在下一次科技创新和产业变革的浪潮中走在最前沿？本文提出以下4个方面具体建议。

（1）要有具有高度前瞻性的顶层设计。顶层设计必须把基础设施建设放在第1位。基础设施建设周期长、任务重、困难大，但从长远发展的角度来说，它的重要性毋庸置疑。过去的几年里，我们目睹一些领域长期的表面繁荣在一夜之间被打回原型的例子，这与先进国家相比呈现出巨大差距。究其原因，都是因为没有在基础设施上下足够的功夫。

（2）要有理性的资源分配机制。要让有能力、有动力、真正活跃在一线的科研人员得到他们应该得到的资源，非理性的资源分配体系所造成的负面影响不仅仅是资源的浪费，更是不正学风的根本原因。要彻底打破靠资历、靠宣传、靠关系和“分蛋糕”的资源分配体系。

（3）要积极推进开放和合作共赢的理念。科学研究本来就是所有科研人员共同的事业。在AI for Science的新框架下，“自给自足、小农作坊”的研究模式将难以适合未来发展的需求。只有合作共赢，才能充分调动科研人员的潜力和积极性，加快提升整体科研创新的能力。

（4）要加强学术风气的建设。学术风气是决定中

① DeepModeling. [2023-12-27]. https://github.com/deepmodeling.

国科技创新能不能成功的最重要的因素之一，也是决定AI for Science在中国能不能顺利发展的最重要的因素之一。要积极鼓励年轻人提出新思想、新观念，鼓励对各种学术观点的质疑和挑战，积极倡导实事求是、有一说一的风气。让学术会议和学术讨论回归其本来的目标。让一些专注于搞虚假宣传、在领导面前画大饼的风气在中国失去生存的空间。

希望我国科学家珍惜目前AI for Science的良好发展势头，紧密合作，紧紧抓住AI for Science这个千载难逢的机会，争取在下一轮的科技创新浪潮中走在前沿，为人类的科技发展作出应有的贡献。

## 参考文献


1 Dirac P A M. Quantum mechanics of many-electron systems. Proceedings of the Royal Society A, 1929, 123(792): 714-733.

2 E W N, Ma C, Wojtowytsch S, et al. Towards a mathematical understanding of neural network-based machine Learning: What we know and what we don't. 2020, doi: arXiv: 2009.10713v3.

3 Jumper J, Evans R, Pritzel A, et al. Highly accurate protein structure prediction with AlphaFold. Nature, 2021, 596: 583-589.

4 Zhang L F, Han J Q, Wang H, et al. End-to-end symmetry preserving inter-atomic potential energy model for finite and extended systems// Proceedings of the 32nd International Conference on Neural Information Processing Systems. New York: Curran And Associates, Inc., 2018: 4441-4451.

5 Zhang L F, Wang H, E W N. Reinforced dynamics for the enhanced sampling in large atomic and molecular systems. The Journal of Chemical Physics, 2018, 148: 124113.

6 E W N, Han J Q, Zhang L F. Machine learning-assisted modeling. Physics Today, 2021, 74(7): 36-41.

7 Guo Z Q, Lu D H, Yan Y J, et al. Extending the limit of molecular dynamics with ab initio accuracy to 10 billion atoms// Proceedings of the 27th ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming. Seoul: ACM, 2022: 205-218.

8 Zeng J Z, Zhang D, Lu D H, et al. DeePMD-kit v2: A software package for deep potential models. Journal of Chemical Physics, 2023, 159(5): 054801.

# AI helps to establish a new paradigm for scientific research

E Weinan
（1 Perking University, Beijing 100871, China;
2 AI for Science Institute, Beijing, Beijing 100084, China）

**Abstract** The main purpose of scientific research is to discover fundamental principles and solve practical problems. Although tremendous progress has been made on both fronts, the lack of effective tools and efficient organizational structure still stands as the main bottleneck for scientific progress. The rapid development of artificial intelligence (AI) offers a new possibility. In recent years, deep learning has had an impressive performance, both in helping to solve fundamental scientific problems and in improving the effectiveness of scientific research tools. A new set of infrastructure is emerging, leading us to a new paradigm, the "Android paradigm", for doing scientific research.

**Keywords** scientific research driven by AI, scientific computing, Android paradigm

**鄂维南** 中国科学院院士。北京大学讲席教授，北京科学智能研究院院长。主要从事计算数学、应用数学，机器学习及其在力学、物理、化学和工程等领域中的应用等方面的研究。E-mail: weinan@math.pku.edu.cn

**E Weinan** Academician of Chinese Academy of Sciences. Professor of Peking University, Director of AI for Science Institute, Beijing. His main research interest is computational science, applied mathematics, machine learning and applications in mechanics, physics, chemistry and engineering. E-mail: weinan@math.pku.edu.cn

■责任编辑：文彦杰